Human tracking with multiple parallel metrics

P. M. Birch*, W. Hassan, R. C. D. Young, C.R. Chatwin

*p.m.birch@sussex.ac.uk, Dept. of Engineering and Design, University of Sussex, Falmer, UK, BN1 9QT

Keywords: HOG, Correlation, Tracking

Abstract

The tracking of humans in a video stream has become one of the most desirable computer vision tasks over the past few years. It remains however a difficult problem and the reliability of systems is often dependent on getting good lighting and clear video images. This paper reports on the development of our PHACT tracker: parallel HOG and correlation tracking. This system uses a cascade of tracking algorithms that enhances the reliability and robustness of the system as a whole when used in difficult conditions.

1 Introduction

The tracking of humans in video remains a difficult problem. People move in a nonlinear and unpredictable manner, are non rigid, and there is a wide degree of variation between different people. There are many applications for such a system including security alerts, complete localisation of several people [1], counting people in groups [2] and behaviour analysis [3].

Tracking can be broken down into two parts: the descriptor, which gives a metric of how likely the pixels belong to the object of interest, and the predictor, which keeps a track of the descriptor over the multiple frames. An example predictor is the Kalman filter and the descriptor could be a colour histogram.

There are a number of problems that must be addressed in visual tracking:

- Robustness - the tracker must cope with changing view angle, changing lighting, image clutter and noise.
- Adaptivity - due to the moving camera and object, the shape of the object will change. The tracker must cope with this.
- The drifting problem[4] - if the tracker's template is updated automatically, background pixels may enter the training set and cause the tracker to lock on to false targets.
- Real time - the tracker must work at a reasonable frame rate on conventional hardware.

It is the aim of this paper to track all humans within an image on standard CPU architectures; thus the tracking must be lightweight. We also assume the video is of low quality, noisy and possibly out of focus. There are a large number of tracking methods within the scientific literature; many are, however, concerned with the tracking of large objects in well illuminated stable environments [5] [6] and so they are not applicable. SIFT and SURF based trackers are examples of this and these require sharply imaged objects that consist of many hundreds of pixels. For tracking smaller objects intensity histograms and colour (hue) based trackers are effective but lack spatial information about the object, making them prone to locking onto the wrong targets. They are, however, fast to calculate.

Correlation based tracking has had a great degree of success for target tracking and identification. Much of the research in recent years has focussed on composite filters that combine multiple input training images, image noise, clutter structure, and out-of-class training images to produce a robust filter. These filters are robust to noise, intensity variations and can work in real time. Alone, the filter has no adaptivity.

The predictive component of the tracker ensures that the track remains on the correct object and is

especially important when two objects cross each other. The simplest predictor is to look for overlap between frames or find the nearest object. A more intelligent approach is to measure the object's velocity vector. Kalman filters are the classic method for achieving this. However, these are linear so they have been extended to the non-linear Extended Kalman filter and unscented Kalman [7]. Particle filters are also widely used [8]. They cope well with the changing direction of the objects but are best suited for extended objects within the image and can suffer from sampling problems. To overcome the problems encountered by the target changing direction most state of the art trackers (e.g. [6], [9]) now use exhaustive search based methods, i.e., apply the descriptor to all reasonable possible locations. This paper has opted for this approach.

Several groups have attempted this task. Some of the first methods used techniques such as frame differencing, motion and colour to detect humans [10].

There have been a number of papers that look at the tracking problem alone and leave the detection to a human operator. Hassan [8] used particle filters, colour histograms, and optical flow. Yilmaz [9] and Yang [11] provide a survey of many of the major techniques. This paper is concerned with both detection and tracking.

For the detection of humans in images there are a number of template matching schemes that have been developed. One of the most successful is the use of Haar wavelets [12] to perform face detection. Although fast to compute, the results can be limited. The use of histograms of orientated gradients (HOG) has proved to be robust [13] and is used in numerous applications.

Dalal's HOG technique[13] uses a grid of edge detector gradients. A histogram of the orientations of the edges is produced from the training set and this data is fed into a linear support vector machine (SVM). To locate a possible target the HOG descriptors are again calculated and a sliding window is tested against the SVM classifier to determine if the window contains the object. Strictly speaking the HOG method is not a full tracker. It is purely a detector.

PROST [6] uses the track by detection method to track generic objects. It uses three cascaded detectors: optical flow, an online random forest and correlation. This is an attempt to balance plasticity with robustness. The optical flow is very plastic but it will soon suffer from the drifting problem. The correlator is very robust but it does not update at all, and the online random forest is somewhere between the two. PROST then tracks through a cascade of these three trackers. If the correlator finds the object, it uses this result. If it does not find the object it will use the random forest and if that fails it will use the optical flow. This produces a very robust tracker although it is designed to track rigid bodies rather than humans.

This paper develops this idea further and presents the PHACT: parallel HOG and correlation filter which has been designed to track humans. It differs in some key aspects to PROST. Firstly, a HOG detector is used to locate all possible humans within the image and, secondly, the correlator is no longer a fixed template, but an adaptive system.

2 PHACT design

Unlike many of the systems described in the introduction, which use a track by detection design, we have opted for a new type of tracker that we call *track by class*. The basic design philosophy is that the algorithm locates all the possible objects of the class within the image and tracks them. The objects can be differentiated by set of correlation filters. The algorithm thus consists of three components:

- Object classification
- Region of support extraction
- Object detection through cross correlation

The object classification is performed using a HOG based classifier. The classifier has to be to trained off-line for the specific set of objects. The tracker is therefore not suitable for tracking any arbitrary object without a training period. However, the trained classes can be rather generic such as people or vehicles, making the method suitable for crime detection applications.

Once the set of objects are detected in the image frame, a rectangular set of coordinates for each object is returned. This then feeds into a correlation

algorithm. The correlation peak is then detected and this is used as the final track result.

This correlation mask is used in subsequent frames until the HOG detector again finds a suitable target and the mask is replaced with the new image.

Occasionally the HOG detector will produce a false positive: it could for example lock on to an area of road. This will then be fed to the correlator which will then produce a very good match since it is correlating the same two images with each other. Without suppressing this, the PHACT would permanently lock onto the background. The algorithm overcomes this by comparing the HOG rectangle with a running average background image. If the HOG rectangle correlates more strongly with the background than the current frame, the track is rejected.

3 Correlation Filter Design

Several designs of correlator have been tested. The simplest is the normalised cross correlation. This can be further improved by band limiting the image by DOG filtering the templates [14]. Both of these options only work on the single previous state. The tracker can be improved further by comparing several past templates. If the nth past template is described as T_n we could test each template individually for all n:

$$C = \sum_{n=1}^{N} I \star T_n \tag{1}$$

where $\star$ is the cross-correlation operator and I is the input image. We would then look for a peak in C. This is rather computationally intensive but we note since the correlation operator is linear:

$$C = I \star \sum_{n=1}^{N} T_n \tag{2}$$

The problem is now that there is probably a large degree of similarity between individual filters since they are from the same object, meaning that I will actually correlate against a number of the filters T_n making the value of C rather unstable for different inputs. To overcome this we can replace the multiple set of T_n with a single filter that encompasses all the individual templates and has a number of design criteria added in. This is known as a composite filter. There are a number different designs but we have incorporated the optimum trade off maximum average correlation height (OT-MACH) filter [15] due to its known performance.

The filter works by attempting maximise the average correlation height (ACH) for all the templates. It attempts to minimise the average correlation energy (ACE), the average similarity matrix (ASM) and output noise variance (ONV) of the filter.

The ASM is a measure of how similar each correlation template is to the others. By minimising it, the filter then gives the same output correlation value no matter which template the input actually matches against. Minimising the ACE forces the filter to give a sharp peak when a match is produced. The ONV is a measure of the filter's ability to reject noise and clutter.

The filter in frequency space is then given by [15]

$$h = D^{-1} m^* \tag{3}$$

where

$$D = \alpha P + \beta D_x + \gamma S_x \tag{4}$$

where P is the noise power spectral density, D_x is the mean power spectral density of the templates, and S_x is the absolute mean difference between the mean Fourier transform of the templates and each template, ie the variance of the Fourier transform of the templates. m^* is the complex conjugate of the mean of the Fourier transform of the template images, T_n. D is a two dimensional array so the $^{-1}$ operator is a pixel level divide, rather than an array inversion (i.e. equivalent to a Matlab $./$).

α, β, γ are tuning parameters that allow the adjustments of the discrimination of the filter and its noise rejection ability. Five past states, as produced by the HOG filter, where used to train the MACH. It is the output of this filter, i.e., the position of the peak in the correlation output, that is used as the track result. One advantage the MACH filters have over single template filters is that since the filter is trained on multiple angles and multiple scales, a degree of out of plane rotation and scale invariance is introduced. To perform the correlation, the inverse Fourier transform is calculated of h and this cross-correlated in the space domain with the current frame.

4 Results

The algorithm was run on an example video that contains one person that is constantly changing speed and direction. A screen shot is shown in Figure 1. The HOG detector found the subject 26% of the time. With the PHACT tracker person was detected 99% of the time. Figure 2 shows the PHACT working on a crowded street [16].

The MACH filter has a large degree of invariance to image degradation and lighting changes. This is demonstrated below. A video sequence was recorded with a fixed exposure time and aperture whilst the lights in the room where changed (see Figure 3. The MACH filter can still determine the position of the person, whilst other techniques such as the colour histogram fail. The filter is also extremely robust against noise as shown in Figure 4 where 30% salt and pepper noise has been added to the image.

It can been seen from the results that the HOG clearly performs badly with the test video. The addition of the matched filter correlation improves this, and it is further improved with the addition of the DOG filtering and MACH filtering. The HOG filter finds most people and again the correlator corrects for any errors.

Figure 1. Example video sequence with HOG detection (blue rectangle) and PHACT tracker (red star)

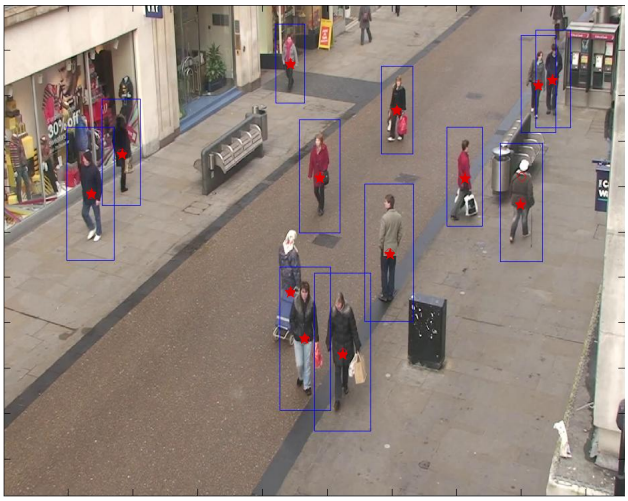

Figure 2. Example images of the PHACT correlation filter working on a busy street. The blue boxes indicate a HOG person detection. Note that the woman with the wheeled shopping bag (centre bottom) is partially occluded. The HOG track (blue squares) has failed but the correlator continues the track (red star).

5 Conclusion

We have demonstrated a new concept for a tracking algorithm, PHACT, which builds upon the HOG detector and incorporates a composite correlation filter. The tracker has a good success rate and it has been shown to be resistant to noise, clutter, lighting and colour changes. The design overcomes some of the problems with correlation tracking, namely the lack of any adaptability. By using the HOG detector to provide possible candidates the drift problem is also removed since we will always be updating the correlator with a human target.

References

[1] S.-I. Yu, Y. Yang, and A. Hauptmann, "Harry Potter's Marauder's Map: Localizing and Tracking Multiple Persons-of-Interest by Nonnegative Discretization," *IEEE CVPR*, 2013.

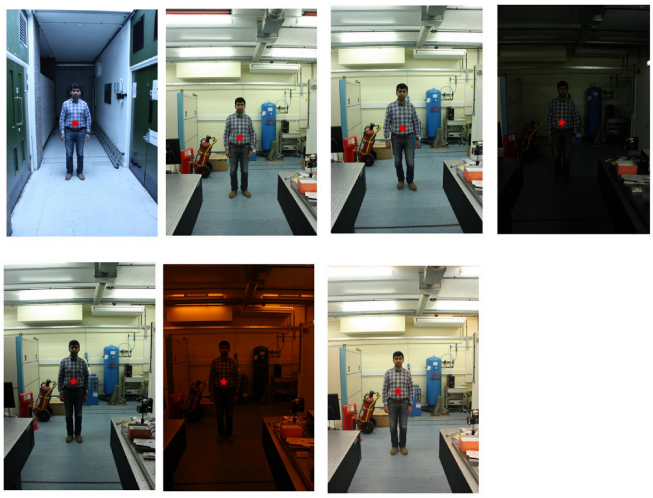

Figure 3. Example images of the PHACT correlation filter working in different lighting

Figure 4. Example images of the PHACT correlation filter working with noise (the image has been degraded with 30% salt and pepper noise).

[2] D. Fehr, R. Sivalingam, V. Morellas, N. Papanikolopoulos, O. Lotfallah, and Y. Park, "Counting People in Groups," *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, pp. 152–157, 2009.

[3] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision And Image Understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[4] L. Matthews, T. Ishikawa, and S. Baker, "The template update problem," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 6, pp. 810–815, 2004.

[5] M. Chen, S. K. Pang, T. J. Cham, and A. Goh, "Visual tracking with generative template model based on riemannian manifold of covariances," *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp. 1–8, 2011.

[6] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, "PROST: Parallel robust online simple tracking," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 723–730, 2010.

[7] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," *Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pp. 153–158, 2000.

[8] W. Hassan, N. Bangalore, P. Birch, R. Young, and C. Chatwin, "An adaptive sample count particle filter," *Computer Vision And Image Understanding*, vol. 116, pp. 1208–1222, Dec. 2012.

[9] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm Computing Surveys*, vol. 38, pp. 13–es, Dec. 2006.

[10] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *Pattern Analysis and*

Machine Intelligence, IEEE Transactions on, vol. 19, no. 7, pp. 780–785, 1997.

[11] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomputing*, vol. 74, pp. 3823–3831, Nov. 2011.

[12] P. Viola and M. Jones, "Robust real-time face detection," *International Journal Of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, p. 886, 2005.

[14] L. Jamal-Aldin, R. Young, and C. Chatwin, "Synthetic discriminant function filter employing nonlinear space-domain preprocessing on bandpass-filtered images," *Applied Optics*, vol. 37, no. 11, pp. 2051–2062, 1998.

[15] A. Mahalanobis, B. Kumar, S. Song, S. Sims, and J. Epperson, "Unconstrained Correlation Filters," *Applied Optics*, vol. 33, no. 17, pp. 3751–3759, 1994.

[16] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *CVPR*, pp. 3457–3464, June 2011.